Recap

P-values based on multi sample splitting

need to avoid "double dipping" using the data twice for variable selection and using statistical inference (tests, confidence intervals) afterwards

$\rightsquigarrow$ sample splitting

multiple sample splitting is much more reliable and statistically better than splitting once

Fixed design linear model

$$Y = X\beta^0 + \varepsilon$$

split the sample into two parts I_1 and I_2 of equal size $\lfloor n/2 \rfloor$

- use (e.g.) Lasso to select variables based on I_1: $\hat{S}(I_1)$
- perform low-dimensional statistical inference on I_2 based on data $(X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2})$;
  for example using the t-test for single coefficients β_j^0

  due to independence of I_1 and I_2, this is a "valid" strategy (see later)

Validity of the (single) data splitting procedure

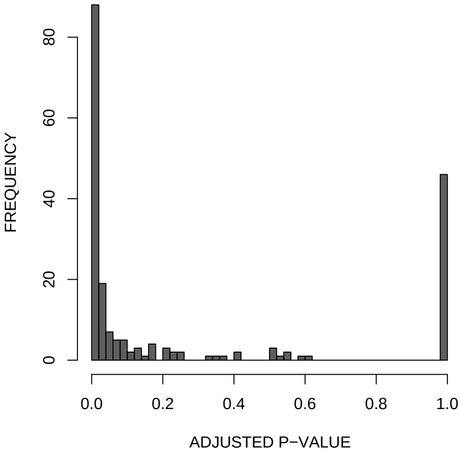
consider testing $H_{0,j}: \ \beta_j^0 = 0$ versus $H_{A,j}: \ \beta_j^0 \neq 0$

assume Gaussian errors for the fixed design linear model:
thus, use the t-test on the second half of the sample I_2 to get a p-value

$$P_{\mathrm{raw},j} \text{ from } t\text{-test based on } X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2}$$

$P_{\mathrm{raw},j}$ is a valid p-value (controlling type I error) for testing $H_{0,j}$ if $\hat{S}(I_1) \supseteq S_0$, i.e., the screening property holds a p-value lottery depending on the random split of the data

motif regression $n = 287$, $p = 195$



$\rightsquigarrow$ should aggregate/average over multiple splits!

Multiple testing and aggregation of p-values

the issue of multiple testing:

$$\tilde{P}_j = \begin{cases} P_{\mathrm{raw},j} \text{ based on } Y_{I_2}, X_{I_2}^{(\hat{S}(I_1))} & \text{, if } j \in \hat{S}(I_1), \\ 1 & \text{, if } j \notin \hat{S}(I_1) \end{cases}$$

thus, we can have at most $|\hat{S}(I_1)|$ false positives
$\leadsto$ can correct with Bonferroni with factor $|\hat{S}(I_1)|$ (instead of factor p) to control the familywise error rate

$$\tilde{P}_{\mathrm{corr},j} = \min(\tilde{P}_j \cdot |\hat{S}(I_1)|, 1) \ (j = 1, \ldots, p)$$

decision rule: reject $H_{0,j}$ if and only if $\tilde{P}_{\mathrm{corr},j} \le \alpha$
$\leadsto$ FWER $\le \alpha$